Tagset Design and Inflected Languages

David Elworthy
Sharp Laboratories of Europe Ltd
Edmund Halley Road
Oxford Science Park
Oxford OX4 4GA
dahe@sharp.co.uk

Abstract

An experiment designed to explore the relationship between tagging accuracy and the nature of the tagset is described, using corpora in English, French and Swedish. In particular, the question of internal versus external criteria for tagset design is considered, with the general conclusion that external (linguistic) criteria should be followed. Some problems associated with tagging unknown words in inflected languages are briefly considered.

1 Tagset Design

Tagging by means of a Hidden Markov Model (HMM) is widely recognised as an effective technique for assigning parts of speech to a corpus in a robust and efficient manner. An attractive feature of the technique is that the algorithm itself is independent of the (natural) language to which it is applied. All of the "knowledge engineering" is localised in the choice of tagset and the method of training. Typically, training makes use of a manually tagged corpus, or an untagged corpus with some initial bootstrapping probabilities. Some attention has been given to how to make such techniques effective; for example Cutting et al. (1992) suggest ways of training trigram taggers, and Merialdo (1994) and Elworthy (1994) consider the amount and quality of the seeding data needed to construct an accurate tagger.

In training a tagger for a given language, a major part of the knowledge engineering required can therefore be localised in the choice of the tagset. The design of an appropriate tagset is subject to both external and internal criteria. The external criterion is that the tagset must be capable of making the linguistic (for example, syntactic or morphological) distinctions required in the output corpora. Tagsets used in the past have included varying amounts of detail. For example, the Penn treebank tagset (Marcus et al., 1993) omits a number of the distinctions which are made in the LOB and Brown tagsets on which it is based (Garside et al., 1987; Francis and Kučera, 1992) in cases where the surface form of the words allows the distinctions to be recovered if they are needed. Thus, the auxiliary verbs *be*, *do* and *have* have the same tags as other verbs in Penn, but are each separated out in the LOB tagset.

A second design criterion on tagsets is the internal one of making the tagging as effective as possible. As an example, one of the most common errors made by taggers with the LOB and Brown tagsets is mistagging a word as a subordinating conjunction (CS) rather than a preposition (IN), or vice-versa (Macklovitch, 1992). A higher level of syntactic analysis indicating the phrasal structure would be required to predict which tag is correct, and this information is not available to fixed-context taggers. The Penn treebank therefore uses a single tag for both cases, leaving the resolution – if required – to some other process. Similarly, most tagsets do not distinguish transitive and intransitive verbs, since taggers which use a context of only two or three words will generally not be able to make the right predictions. Distinctions of this sort are usually found only in corpora such as Susanne which are parsed as well as tagged.

The problem of tagset design becomes particularly important for highly inflected languages, such as Greek or Hungarian. If all of the syntactic variations which are realised in the inflectional system were represented in the tagset, there would be a huge number of tags, and it would be practically impossible to implement or train a simple tagger. Note in passing that this may not as serious a problem as it first appears. If the language is very highly inflected, it may be be possible to do all (or a large part) of the work of a tagger with a word-by-word

morphological analysis instead. Nevertheless, there are many languages which have enough ambiguity that tagging is useful, but a rich enough tagset that the criteria on which it is designed must be given careful consideration.

In this paper, I report two experiments which address the *internal* design criterion, by looking at how tagging accuracy varies as the tagset is modified, in English, French and Swedish. Although the choice of language was dictated by the corpora which were available, they represent three different degrees of complexity in their inflectional systems. English has a very limited system, marking little more than plurality on nouns and a restricted range of verb properties. French has a little more complexity, with gender, number and person marked, while Swedish has more detailed marking for gender, number, definiteness and case. As a subsidiary issue, we will also look at how the tagger performs on unknown words, i.e. ones not seen in the training data. The usual approach here is to hypothesise all tags in the tagset for an unknown word, other than ones where all the words that may have the tag can be enumerated in advance (closed class tags). HMM taggers often perform poorly on unknown words.

Alternative tagsets were derived by taking the initial tagset for each corpus (from manual tagging of the corpus) and condensing sets of tags which represent a grammatical distinction such as gender into single tags. The changes were then applied to the training corpus. This allows us to effectively produce a corpus tagged according to a different scheme without having to manually re-tag the corpus. The changes in the tagsets were motivated purely by grammatical considerations, and did not take the errors actually observed into account. In general what we will look at in the results is how the tagging accuracy changes as the *size* of the tagset changes. This is a deliberately naive approach, and it is adopted with the goal of continuing in the relatively "knowledge-free" tradition of work in HMM tagging. The aim of the experiment is to determine, crudely, whether a bigger tagset is better than a smaller one, or whether external criteria requiring human intervention should be used to choose the best tagset. The results for the three languages turn out to be quite different, and the general conclusion (which is the overall contribution of the paper) will be that the external criterion should be the one to dominate tagset design: there is a limit to how knowledge-free we can be.

As a preliminary to this work, note that it is hard to reason about the effect of changing the tagset. It can be argued that a smaller tagset should improve tagging accuracy, since it puts less of a burden on the tagger to make fine distinctions. In information-theoretic terms, the number of decisions required is smaller, and hence the tagger need contribute less information to make the decisions. A smaller tagset may also mean that more words have only one possible tag and so can be handled trivially. Conversely, more detail in the tagset may help the tagger when the properties of two adjacent words give support to the choice of tag for both of them; that is, the transitions between tags contribute the information the tagger needs. For example, if determiners and nouns are marked for number, then the tagger can effectively model agreement in simple noun phrases, by having a higher probability for a singular determiner followed by a singular noun that it does for a singular determiner followed by a plural noun. Theory on its own does not help much in deciding which point of view should dominate.

2 The experiments

2.1 Design of the experiments

Two experiments were conducted on three corpora: 300k words of Swedish text from the ECI Multilingual CD-ROM, and 100k words each of English and French from a corpus of International Telecommunications Union text[1]. In the first experiment the whole of each corpus was used to train the model, and a small sample from the same text was used as test data. For the second experiment, 95% of the corpus was used in training and the remainder in testing. The importance of the second test is that it includes unknown words, which are difficult to tag. The tagsets were progressively modified, by textually substituting simplified tags for the original ones and e e-running the training and test procedures using the modified corpora. The changes to the tagset are listed below. In the results that follow, we will identify tagset that include a given distinction with an uppercase letter and ones that do not with a lowercase letter; for example G for a tagset that marks gender, and g for one that does not.

Swedish The changes made were entirely based on inflections.

- **G** Gender: masculine, neuter, common gender ("UTR" in the tagset).
- **N** Number: singular, plural.
- **D** Definiteness: definite, indefinite.
- **C** Case: nominative, genitive.

[1]The English and French corpora were kindly supplied to us by Tony McEnery, and are translation-equivalent. See McEnery et al. (1994) for details.

French The changes other than *V* were based on inflections.

- **G** Gender: masculine, feminine.
- **N** Number: singular, plural.
- **P** Person: identified as 1st to 6th in the tagset.
- **V** Verbs: treat *avoir* and *etre* as being the same as any other verb.

English The changes here are more varied than for the other languages, and generally consisted of removing some of the finer subdivisions of the major classes. The grouping of some of these changes is admittedly a little *ad hoc*, and was intended to give a good distribution of tagset sizes; not all combinations were tried.

- **C** Reduce specific conjunction classes to a common class, and simplify one adjective class.
- **A** Simplify noun and adverb classes.
- **P** Simplify pronoun classes.
- **N** Number: all singular/plural distinctions removed.
- **V** Use the same class for *have*, *do* and *be* as for other verbs.

The sizes of the resulting tagsets and the degree of ambiguity in the corpora which resulted appear below. Accuracy figures quoted here are for ambiguous and unknown words only, and therefore factor out effects due to the varying degree of ambiguity as the tagset changes. In fact, this is a rather approximate way of accounting for ambiguity, since it does not take the length of ambiguous sequences into account, and the accuracy is likely to deteriorate more on long sequences of ambiguous words than on short ones.

The tests were run using Good-Turing correction to the probability estimates; that is, rather than estimating the probability of the transition from a tag i to a tag j as the count of transition from i to j in the training corpus divided by the total frequency of tag i, one was added to the count of all transitions, and the total tag frequencies adjusted correspondingly. The purpose in using this correction is to correct for corpora which might not provide enough training data. On the largest tagsets, the correction was found to give a very slight reduction in the accuracy for Swedish, and to improve the French and English accuracies by about 1.5%, suggesting that it is indeed needed.

2.2 Results

The first experiment, with no unknown words, gave accuracies on ambiguous words of 91–93% for Swedish, 94–97% for French and 85–90% for English. The results for English are surprisingly low (for example, on the Penn treebank, the tagger gives an accuracy of 95–96%), and may be due to long sequences of ambiguous words. The results appear in table 1. The figures include the degree of ambiguity, that is, the number of words in the corpus for which more than one tag was hypothesised. The accuracy is plotted against the size of the tagset in figures 1–3, where the numbers on the points correspond to the index of tagsets listed. Summarising the patterns:

Swedish Larger tagset generally gives higher accuracy. The results are quite widely spread.

French Clustered, with an accuracy on all tagsets which do not mark gender of around 96%–96.5%; when gender is marked 94%–94.5%.

English Larger tagset tends to give larger accuracy, though with less of a spread than for Swedish.

The sizes of the tagsets ranged from approximately 80–200 tags for Swedish, 35–90 for French, and 70–160 for English. As discussed above, it is not clear what would happen with larger tagsets, but some experiments based on the Susanne corpus and using tagsets ranging from 236 to 425 tags suggest that the trend to higher accuracy continues with even bigger tagsets.

In the second experiment, the test corpora included "unknown" words, which had not been seen during training, and for which the tagger hypothesises all open-class tags. Two results are interesting to look at here: the accuracy on the unknown words, and the accuracy on words which were ambiguous but were found in the training corpus. The results, in outline, are:

Swedish Similar results on known words to first experiment. For unknown words, smaller tagsets give higher accuracy.

French For ambiguous words, the pattern and accuracy were similar to first experiment. For unknown words, the pattern of accuracies was again similar, with tagsets that do not include gender giving accuracies of 51%–52%, and those which do giving 45%–46%.

English Ambiguous words gave similar results to the first test. Unknown words show a weak tendency to give higher accuracy on smaller tagsets.

Typical accuracies on ambiguous words were 90–92%, 93–97% and 83–88% for Swedish, French and

Table 1: Results for test with no unknown words

Language	Index	Tagset	Size of tagset	Degree of ambiguity (%)	Ambiguous word accuracy (%)
Swedish	1	GNDC	194	41.57	92.02
	2	GnDC	170	39.29	92.23
	3	GNDc	167	41.49	91.92
	4	GNdC	162	41.45	91.67
	5	gNDC	152	41.54	91.88
	6	GnDc	147	39.21	92.04
	7	GndC	141	37.43	91.86
	8	GNdc	140	41.37	91.63
	9	gNDc	134	41.47	91.82
	10	gNdC	126	41.42	91.34
	11	Gndc	123	37.36	91.74
	12	gnDC	121	39.18	91.35
	13	gNdc	113	41.34	91.29
	14	gnDc	105	39.11	91.28
	15	gndC	96	37.32	91.56
	16	gndc	86	37.25	91.52
French	1	GNPV	87	49.77	94.43
	2	GNPv	80	49.75	94.35
	3	GNpV	76	49.77	94.31
	4	GNpv	74	49.75	94.39
	5	GnPV	64	49.49	94.28
	6	gNPV	62	47.48	96.34
	7	GnPv	57	49.47	94.36
	8	gNPv	55	47.64	96.22
	9	GnpV	53	49.49	94.25
	10	gNpV	51	47.48	96.14
	11	Gnpv	51	49.47	94.36
	12	gnPV	49	47.03	96.34
	13	gNpv	49	47.46	96.10
	14	gnPv	42	47.01	96.30
	15	gnpV	38	47.03	96.30
	16	gnpv	36	47.01	96.34
English	1	CAPNV	153	47.95	89.56
	2	CApNV	150	47.47	89.27
	3	cAPNV	145	47.91	89.50
	4	CAPNv	140	47.95	89.33
	5	CAPnV	137	47.95	89.20
	6	CaPNV	129	47.95	89.20
	7	CAPnv	124	47.95	89.01
	8	capNV	119	47.43	88.94
	9	capnV	108	47.13	88.45
	10	capNv	106	47.43	88.48
	11	capnv	95	47.13	85.42

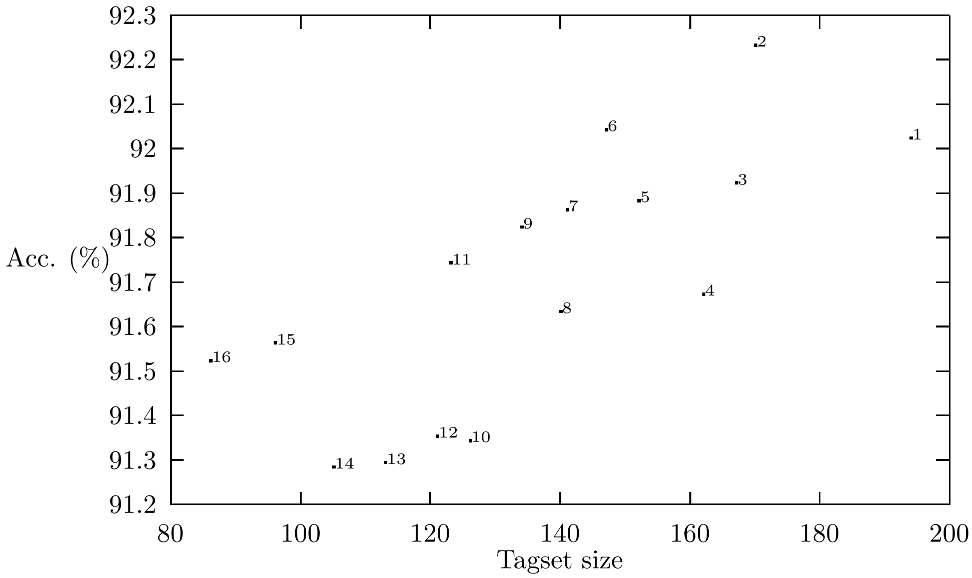

Figure 1: Results for Swedish (no unknown words)

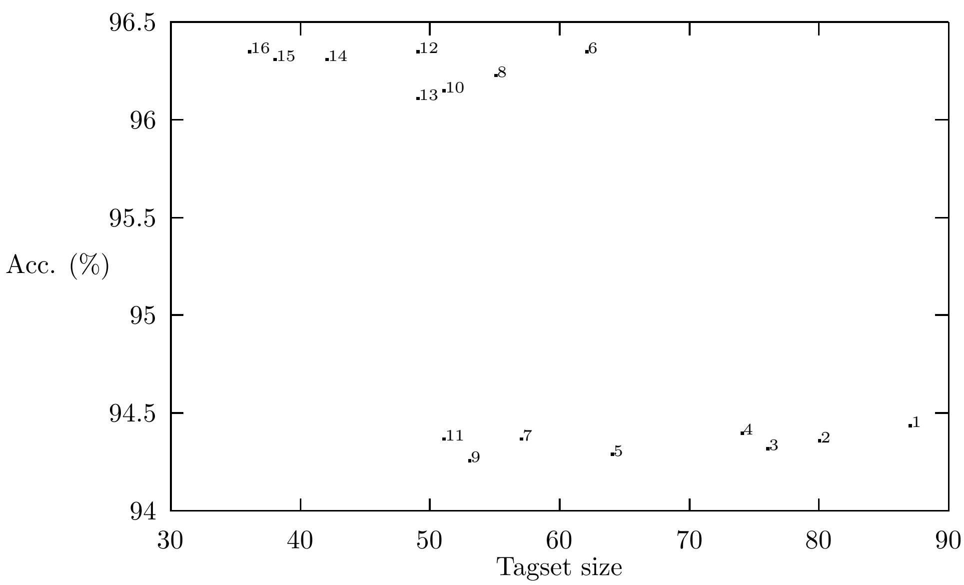

Figure 2: Results for French (no unknown words)

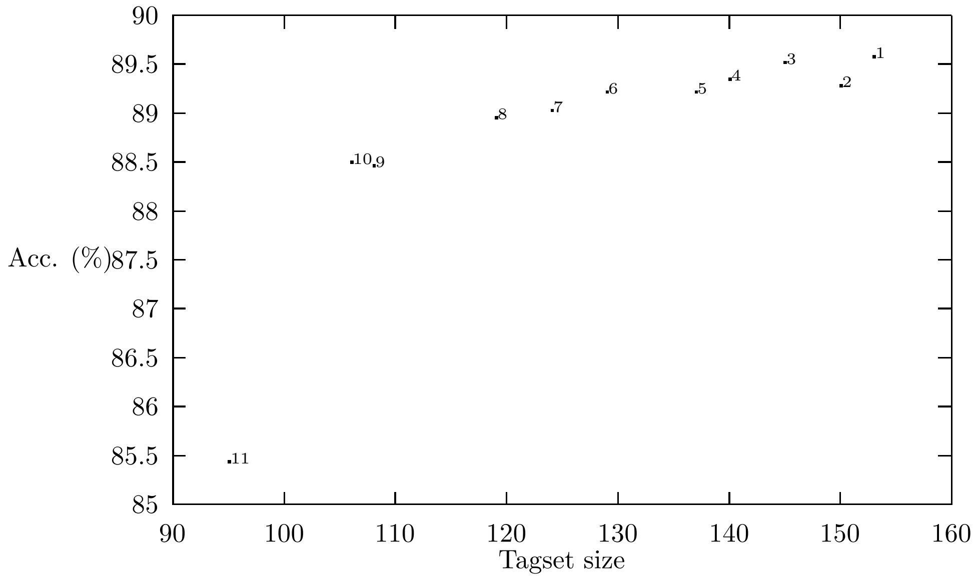

Figure 3: Results for English (no unknown words)

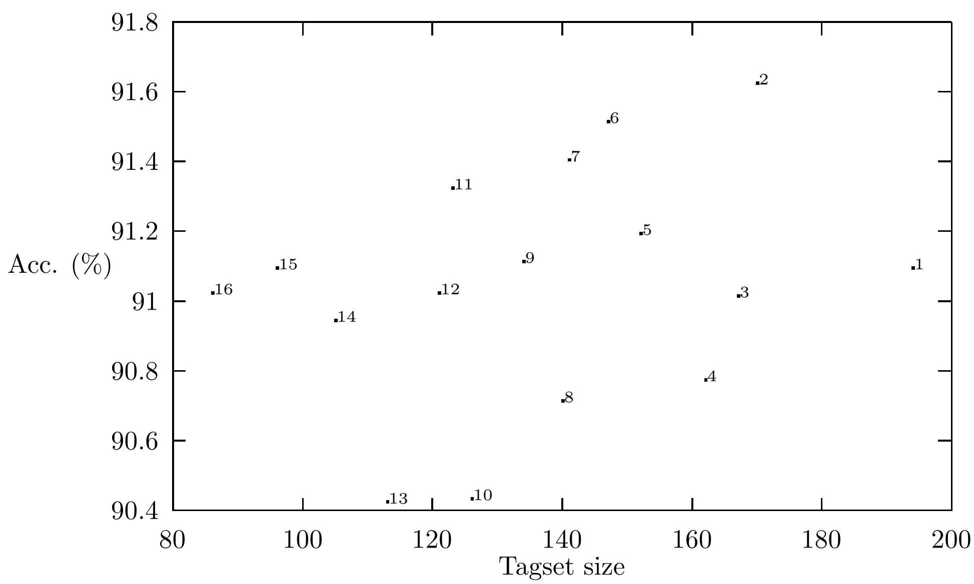

Figure 4: Results for Swedish (with unknown words)

Table 2: Results for test with unknown words

Language	Index	Tagset	Size of tagset	Degree of ambiguity (%)	Ambiguous word accuracy (%)	Unknown word accuracy (%)
Swedish	1	GNDC	194	52.60	91.09	23.42
	2	GnDC	170	50.56	91.62	26.28
	3	GNDc	167	52.59	91.01	24.17
	4	GNdC	162	52.48	90.77	28.48
	5	gNDC	152	52.57	91.19	29.33
	6	GnDc	147	50.55	91.51	26.48
	7	GndC	141	48.86	91.40	36.29
	8	GNdc	140	52.46	90.71	28.63
	9	gNDc	134	52.56	91.11	29.48
	10	gNdC	126	52.45	90.43	36.14
	11	Gndc	123	48.85	91.32	36.44
	12	gnDC	121	50.46	91.02	34.73
	13	gNdc	113	52.43	90.42	36.24
	14	gnDc	105	50.45	90.94	35.39
	15	gndC	96	48.75	91.09	48.00
	16	gndc	86	48.74	91.02	47.85
French	1	GNPV	87	58.37	93.86	45.74
	2	GNPv	80	58.35	93.86	45.41
	3	GNpV	76	58.37	93.78	45.58
	4	GNpv	74	58.35	93.78	45.41
	5	GnPV	64	58.09	93.63	45.74
	6	gNPV	62	56.54	96.50	50.58
	7	GnPv	57	58.07	93.74	46.08
	8	gNPv	55	56.52	96.46	51.25
	9	GnpV	53	58.09	93.75	45.74
	10	gNpV	51	56.54	96.38	50.92
	11	Gnpv	51	58.07	93.78	46.24
	12	gnPV	49	56.09	96.26	50.92
	13	gNpv	49	56.62	96.34	50.75
	14	gnPv	42	56.08	96.26	52.45
	15	gnpV	38	56.09	96.26	52.25
	16	gnpv	36	56.08	96.34	52.59
English	1	CAPNV	153	55.65	87.57	46.49
	2	CAPnv	150	55.17	87.54	46.69
	3	cAPNV	145	55.60	87.46	46.29
	4	CAPNv	140	55.65	87.52	46.09
	5	CAPnV	137	55.65	87.62	51.70
	6	CaPNV	129	55.65	87.43	46.29
	7	CAPnv	124	55.65	87.48	51.70
	8	capNV	119	55.13	87.38	46.49
	9	capnV	108	55.00	83.66	56.51
	10	capNv	106	55.13	83.66	44.29
	11	capnv	95	55.00	83.56	55.11

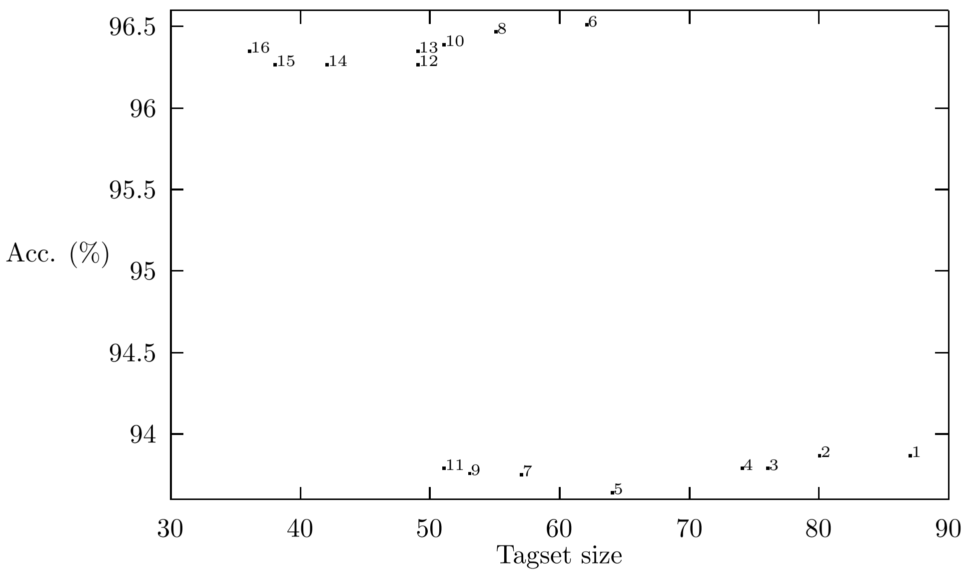

Figure 5: Results for French (with unknown words)

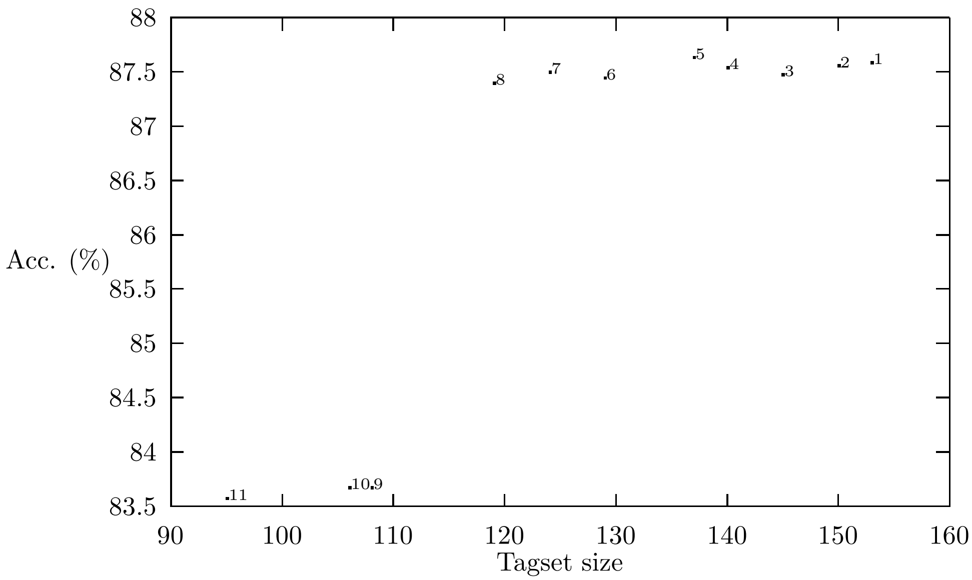

Figure 6: Results for English (with unknown words)

English respectively, with the corresponding accuracies on unknown words being 25–50%, 45–52% and 44–58%. Table 2 lists the results, giving the tagset size, the degree of ambiguity and the accuracies on known ambiguous and unknown words. The ambiguous word accuracy is plotted in figures 4–6.

What seems to come out from these results is that there is not a consistent relationship between the size of the tagsets and the tagging accuracy. The most common pattern was for a larger tagset to give higher accuracy, but there were notable exceptions in French (where gender marking was the key factor), in Swedish unknown words (which show the reverse trend) and in English unknown words (which show no very clear trend at all). This seems to fit quite well with the difficulties that were suggested above in reasoning about the effect of tagset size. The main conclusion of this paper is therefore that the knowledge engineering component of setting up a tagger should concentrate on optimising the tagset for external criteria, and that the internal criterion of tagset size does not show sufficient generality to be taken into account without prior knowledge of properties of the language. Perhaps this is not too surprising, but it is useful to have an experimental confirmation that the linguistics matters rather than the engineering.

3 Unknown words

One final observation about the experiments: the accuracy on unknown words was very low in all of the tests, and was particularly bad in Swedish. The tagger used in the experiments took a very simple-minded approach to unknown words. An alternative that is often used is to limit the possible tags using a simple morphological analysis or some other examination of the surface form of the word. For example, in a variant of the English tagger which was not used in these experiments, a module which reduces the range of possible tags based on testing for only seven surface characteristics such as capitalisation and word endings improved the unknown word accuracy by 15-20%.

The results above show that if it were not for unknown words, there might be some argument for favouring larger tagsets, since they have some tendency to give a higher accuracy. A tentative experiment on the contribution of using morphological or surface analysis in French and Swedish was therefore carried out. Firstly, in both languages, the unknown words from the second experiment were looked up in the lexicon trained from the full corpus to see what tags they might have. For Swedish, 96% of the unknown words came from inflected classes, and had a single tag; for French the figure was about 60%. In both cases, very few of the unknown words (less than 1%) had more than one tag. This provides some hope that an inflectional analysis might should help considerably with unknown words[2]. For confirmation, the list of French unknown words was given to a French grammarian, who predicted that it would be possible to make a good guess at the correct tag from the morphology for around 70% of the words, and could narrow down the possible tags to two or three for about a further 25%. However, further research is needed to determine how realistic these estimates turn out to be.

4 Conclusion

We have shown how a simple experiment in changing the tagset shows that the relationship between tagset size and accuracy is a weak one and is not consistent against languages. This seems to go against the "folklore" of the tagging community, where smaller tagsets are often held to be better for obtaining good accuracy. I have suggested that what is important is to choose the tagset required for the application, rather than to optimise it for the tagger. A follow-up to this work might be to apply similar tests in other languages to provide a further confirmation of the results, and to see if language families which similar characteristics can be identified. A further conclusion might be that when a corpus is being tagged by hand, a large tagset should be used, since it can always be reduced to a smaller one if the application demands it. Perhaps the major factor we have to set against this is the danger of introducing more human errors into the manual tagging process, by increasing the cognitive load on the human annotators.

References

[Cutting et al.1992] Doug Cutting, Julian Kupiec, Jan Pedersen, and Penelope Sibun (1992). A Practical Part-of-Speech Tagger. In *Third Conference on Applied Natural Language Processing. Proceedings of the Conference. Trento, Italy*, pages 133–140, Association for Computational Linguistics.

[Elworthy1994] David Elworthy (1994). Does Baum-Welch Re-estimation Help Taggers? In *Fourth Conference on Applied Natural Language Processing. Proceedings of the Conference. Stuttgart, Germany*, pages 53–58, Association for Computational Linguistics.

[2] Although the figures here are likely to represent a best case, given how little of the corpora was held out.

[Francis and Kučera1992] W. N. Francis and F. Kučera (1992). *Frequency Analysis of English Usage*. Houghton Mifflin.

[Garside et al.1987] Roger Garside, Geoffrey Leech, and Geoffrey Sampson (1987). *The Computational Analysis of English: A Corpus-based Approach*. Longman, London.

[Macklovitch1992] Elliott Macklovitch (1992). Where the Tagger Falters. In *Proceedings of the Fourth Conference on Theoretical and Methodological Issues in Machine Translation*, pages 113–126.

[Marcus et al.1993] Mitchell P. Marcus, Beatrice Santorini, and Mary Ann Marcinkiewicz (1993). Building a Large Annotated Corpus of English: The Penn Treebank. *Computational Linguistics*, 19(2):313–330.

[McEnery et al.1994] A. M. McEnery, M. P. Oakes, R. Garside, J. Hutchinson, and G. N. Leech (1994). The Exploitation of Parallel Corpora in Projects ET10/63 and CRATER. In *International Conference on New Methods in Language Processing. Proceedings of the Conference*, pages 108–115, Centre for Computational Linguistics, UMIST.

[Merialdo1994] Bernard Merialdo (1994). Tagging English Text with a Probabilistic Model. *Computational Linguistics*, 20(2):155–171.